

## NONVOLATILE MEMORY CELL WITH MULTIPLE FLOATING GATES

## FORMED AFTER THE SELECT GATE

Yi Ding

## 5 BACKGROUND OF THE INVENTION

[0001] The present invention relates to nonvolatile memories.

[0002] Figs. 1-4 illustrate a flash memory fabrication process described in van Duuren et al., "Compact poly-CMP embedded flash memory cells for one or two bit storage", Proceedings of NVSMW 2003 (Non-Volatile Semiconductor Memory  
10 Workshop), Monterey, California, pages 73-74. Tunnel oxide 150, polysilicon floating gate 160, inter-poly dielectric 164, control gate 170, and a nitride cap layer 172 are fabricated in a stacked structure ("FG/CG stack"). TEOS spacers 176 are formed on both sides of the stack. Then oxide 130 is grown for the access gate.

[0003] AG (access gate) polysilicon 140 is deposited over the FG/CG stack. See  
15 Fig. 2. Polysilicon 140 is polished by chemical mechanical polishing (CMP), as shown in Fig. 3. Then polysilicon 140 is patterned using resist 173 to define the access gate, as shown in Figs. 3 and 4. Source/drain regions 174 are formed to obtain a one-bit memory cell 102 (Fig. 4).

[0004] As noted in the Duuren et al. article, the length of access gate 140 depends on  
20 the mask alignment, "which could lead to an odd-even word line effect in arrays".

[0005] Fig. 5 shows a two-bit memory cell 110 described in the same article. Two FG/CG stack transistors 110L, 110R share an access gate 140. According to the Duuren et al. article, the cell is fabricated with the same process as cell 102, but cell 110 is fully self-aligned and therefore not sensitive to mask misalignment.

25 [0006] Each bit 110L, 110R can be programmed or erased independently of the other bit. The bit can be programmed by Fowler-Nordheim tunneling (FN) or source side injection (SSI). The Duuren et al. article states that the two bit cell has been studied "with 180 bit arrays in a virtual ground configuration". The read, program (SSI) and erase

voltages bit 110R are shown respectively in Figs. 6, 7 and 8. In the read and program operations (Figs. 6 and 7), the “pass” voltage for the control gate in bit 110L (6.0 V) is high enough to turn on the corresponding FG/CG transistor regardless of the state of its floating gate.

- 5    **[0007]**     Alternative fabrication methods for the two bit cells are desirable.

#### SUMMARY

**[0008]**     This section summarizes some features of the invention. Other features are described in the subsequent sections. The invention is defined by the appended claims which are incorporated into this section by reference.

- 10   **[0009]**     The present invention relates to fabrication of a memory cell having multiple floating gates (such as the cell of Fig. 5, for example). In some embodiments, the access gate is formed before the floating gates. In some embodiments, the memory cell also has control gates (like in Fig. 5), and the access gate is formed before the floating and control gates.

- 15   **[0010]**     Below the term “select gate” is used instead of “access gate”.

**[0011]**     In some embodiments, forming the select gate before the floating gate makes it possible to reduce the select gate width below the minimal photolithographic line width. For example, the select gate, or a mask used to pattern the select gate, can be subjected to a horizontal etch (e.g. isotropic etch) to reduce the select gate width.

- 20   Alternatively, the select gate sidewalls can be oxidized and then the oxide can be removed to reduce the gate length.

- [0012]**     In some embodiments, the memory cell includes an FG/CG stack (such as in Fig. 5), but this stack is formed after the select gate and after the dielectric that separates the stack from the select gate (note dielectric 176 in Fig. 5). The stack is formed by a deposition of the FG and CG layers followed by an etch. The etch does not attack the edges of the FG and CG layers adjacent to the dielectric near the select gate. Therefore, the vertical edge of the FG/CG stack near the select gate is not defined by this etch. This is advantageous because when a vertical edge is defined by the etch of the FG/CG stack, the edge may be non-uniform, with the different edges having different profiles. Also, the edge defined by the stack etch may have protruding “shoulders” which impede formation
- 25
- 30

of subsequent layers. The present invention allows the etch requirements to be relaxed.

[0013] In some embodiments, the cell is completely self-aligned (the floating, control and select gates do not depend on the mask alignment), but the invention is not limited to such embodiments.

5 [0014] In some embodiments, substrate isolation regions are formed in a semiconductor substrate. Each substrate isolation region is a dielectric region protruding above the substrate. Then the select gates are formed. The select gates are part of select gate lines. Each select gate line provides select gates for at least one memory row. Then a floating gate layer (e.g. polysilicon) is deposited. The floating gate layer is etched until  
10 the substrate isolation regions are exposed. In some embodiments, the exposure of the substrate isolation regions serves as an end point for the floating gate layer etch.

[0015] In some embodiments, the memory also has control gates. A control gate layer is deposited over the floating gate layer. The control gate layer protrudes upward over each select gate line. These protrusions are exploited to define the control gates in a self-  
15 aligned manner. The floating gates are then also defined in a self-aligned manner.

[0016] One embodiment of the present invention is a nonvolatile memory cell comprising a conductive floating gate. A dielectric layer overlying the floating gate has a continuous feature that overlies the floating gate and also overlays a sidewall of the select gate. The control gate overlies the continuous feature of the dielectric layer. The  
20 continuous feature of the dielectric layer separates the control gate from the select gate.

[0017] Other features and advantages of the invention are described below. The invention is defined by the appended claims.

#### BRIEF DESCRIPTION OF THE DRAWINGS

[0018] Figs. 1-8 shows vertical cross sections of prior art memory cells and  
25 intermediate structures obtained in prior art fabrication processes.

[0019] Fig. 9 is a circuit diagram of a memory array according to an embodiment of the present invention.

[0020] Fig. 10A is a top view of a memory array according to an embodiment of the present invention.

[0021] Fig. 10B is a perspective view showing some features of the memory of Fig. 10A.

[0022] Figs. 11, 12A, 12B, 13A, 13B, 14A, 14B show vertical cross sections of integrated circuit structures according to embodiments of the present invention.

5 [0023] Fig. 15 is a perspective view of an integrated circuit structure according to an embodiment of the present invention.

[0024] Figs. 16, 17, 18, 19A, 19B, 20A, 20B, 21A, 21B, 22, 23A, 23B, 24, 25, 26A, 26B, 27A, 27B, 28, 29A, 29B, 30A, 30B, 31A-31D show vertical cross sections of integrated circuit structures according to embodiments of the present invention.

10 [0025] Figs. 31E, 32 are top views of integrated circuit structures according to embodiments of the present invention.

#### DESCRIPTION OF SOME EMBODIMENTS

[0026] The embodiments described in this section illustrate but do not limit the invention. The invention is not limited to particular materials, process steps, or  
15 dimensions. The invention is defined by the appended claims.

[0027] One embodiment of the invention will now be described on the example of the memory array of Fig. 9. In this example, the array has 4 rows and 5 columns, but any number of rows and columns can be present. Fig. 10A is a top view of the array. Fig. 10B is a perspective view. Each memory cell 110 may have the same structure as in Fig. 5, but  
20 may also have a different structure (see e.g. Fig. 30A). Each cell 110 has two FG/CG stacks per one select gate 140. Conductive select gate lines 140 and conductive control gate lines 170 run through the memory array in the Y direction (row direction). Each row includes one select gate line 140 and two control gate lines 170. The line 140 provides the select gates for that row of cells. One of the lines 170 provides the control gates for the  
25 bits 110L in that row, and the other line 170 provides the control gates for the bits 110R. Bitlines 180 (marked BL0-BL5 for rows 0-5) run in the X direction (column direction). The bitlines contact the corresponding source/drain regions 174 ("bitline regions") in areas 174C (Fig. 10A) marked with a cross. Floating gates 160 are marked with dashed crosses in Fig. 10A. The floating gates can be completely self-aligned (i.e. defined  
30 independently of photolithographic alignment), as described below.

**[0028]** Substrate isolation trenches 220T run through the array in the X direction. Trenches 220T are filled with dielectric 220 (field isolation). Active areas 222 run through the array between the trenches 220T. Each active area 222 includes active areas of individual cells in one memory column. The active area of each cell consists of the cell's source/drain regions 174 and the P type channel region extending between the regions 174.

**[0029]** In each column, each two consecutive memory cells have their adjacent source/drain regions 174 merged into a single contiguous region (referenced by the same numeral 174). Each such region 174 provides the source/drain regions to only two of the memory cells in each column. In each column 1-4 (each column except the first column and the last column), each source/drain region 174 is connected to a source/drain region 174 of an adjacent column. The connections alternate, e.g. one source/drain region 174 in column 1 is connected to a source/drain region 174 in column 0, the next region 174 in column 1 is connected to region 174 in column 2, the next region 174 in column 1 is connected to region 174 in column 0, and so on. Bitline BL1 (column 1) is connected to those regions 174 of column 1 that are connected to column 0; bitline BL2 is connected to those regions 174 in column 1 that are connected to column 2, and so on. Bitlines BL0 and BL5 are each connected to only one column. In some embodiments, these two bitlines are shorted together.

**[0030]** As shown in Fig. 10A, the source/drain regions 174 of each column are separated from the source/drain regions 174 in the adjacent columns by field isolation regions 220.

**[0031]** Some of the figures below illustrate vertical cross sections of intermediate structures obtained during the memory fabrication. The sectional planes are indicated in Fig. 10A by lines X1-X1', X2-X2', Y1-Y1', and Y2-Y2'. The line X1-X1' runs in the X direction through floating gates 160 (through an active area 222). The line X2-X2' runs in the X direction between the floating gates (through a trench 220T). The line Y1-Y1' runs in the Y direction through a select gate line 140. The line Y2-Y2' runs in the Y direction through a control gate line 170 and floating gates 160.

**[0032]** In one embodiment, the memory is fabricated as follows. Substrate isolation regions 220 are formed in P doped substrate 120 by shallow trench isolation technology ("STI"). See Fig. 11 (cross section Y1-Y1'). Each region 220 is a dielectric region

formed in a trench 220T. Suitable STI processes are described in U.S. patent no. 6,355,524 issued March 12, 2002 to Tuan et al.; U.S. patent application no. 10/262,785 filed October 1, 2002 by Yi Ding; and U.S. patent application no. 10/266,378 filed October 7, 2002 by C. Hsiao, all incorporated herein by reference. Other STI and non-STI processes are also possible. Dielectric 220 is sometimes called "STI oxide" hereinbelow because it is silicon dioxide in some embodiments. The invention is not limited to such embodiments or to silicon integrated circuits.

[0033] Substrate isolation regions are also formed in the memory peripheral area (not shown in Fig. 11). The peripheral area contains circuitry needed to access the memory, and may also contain unrelated circuitry (the memory may be embedded into a larger system).

[0034] As shown in Fig. 11, oxide 220 protrudes above the substrate 120. The protruding portions are shown at 220P. An exemplary thickness of portions 220P is 0.12  $\mu\text{m}$  for a 0.18  $\mu\text{m}$  fabrication process (a process with a 0.18  $\mu\text{m}$  minimum line width). The exemplary dimensions given in this section assume a 0.18  $\mu\text{m}$  fabrication process unless mentioned otherwise.

[0035] Dopant is implanted into substrate 120 to form an N type region 604 underlying the memory array. Dopant is also implanted into the substrate around the array to form a surrounding N type region (not shown) extending from the top surface of substrate 120 down to region 604. These implants create a fully isolated P well 120W for the memory array. Region 604 is not shown in the subsequent drawings, and the P well 120W is shown simply as substrate 120.

[0036] Ion implantation steps ("Vt adjust implants") may be performed into the active areas of substrate 120 to adjust the transistor threshold voltages as needed. One such implant is an N type implant (e.g. arsenic) performed into the array to reduce the threshold voltage of the select gate transistors. This implant creates a counterdoped region 230 at the surface of substrate 120. Region 230 may remain type P, but the net P type dopant concentration in this region is reduced.

[0037] In some embodiments, region 230 becomes N type in this counterdoping step.

[0038] Silicon dioxide 130 (Fig. 12A, cross section Y1-Y1', and Fig. 12B, periphery) is thermally grown on the exposed areas of substrate 120 to provide gate dielectric for the

select gates of the memory array and for the peripheral transistors. An exemplary thickness of oxide 130 in the array area is 120 Å. Generally, the oxide thickness depends on the maximum voltage that the oxide 130 is designed to sustain during the memory operation. Oxide 130 can be nitrided when it is being grown, or after it has been grown,  
5 to impede boron diffusion from floating gates 160 into substrate 120.

[0039] In the example shown in Fig. 12B, the peripheral area includes a high voltage transistor area 512H and a low voltage transistor area 512L. Oxide 130 is grown thermally to a thickness of 60 Å over the entire wafer. This oxide is removed from the low voltage area 512L by a masked etch. The wafer is re-oxidized to re-grow silicon  
10 dioxide in area 512L to a thickness of 60 Å. The oxide thickness in the memory array area and in high voltage area 512H increases from 60 Å to 120 Å during this step.

[0040] Thus, oxide 130 in the array area and oxide 130 in the high voltage peripheral area 512H is formed simultaneously in these two oxidation steps. All of oxide 130 in area 512L and part of the oxide 130 in the array area and area 512H are formed  
15 simultaneously in the second oxidation step.

[0041] As shown in Fig. 13A (cross section Y1-Y1') and Fig. 13B (periphery), intrinsic polysilicon layer 140 is formed over the structure by a conformal deposition process (e.g. low pressure chemical vapor deposition, "LPCVD"). Polysilicon 140 fills the spaces between the oxide protrusions 220P in the memory array area. The top  
20 polysilicon surface is planar because the polysilicon portions deposited on the sidewalls of protrusions 220P meet together.

[0042] Fig. 13B may represent either the low voltage or the high voltage transistor area. In some embodiments, there are more than two peripheral areas with different gate oxide thicknesses, and Fig. 13B may represent any of these areas.

[0043] Polysilicon 140 covers the regions 120i (Fig. 13B) at the interface between substrate 120 and field oxide 220 in the peripheral area. Polysilicon 140 will protect the oxide 220 in this area to prevent formation of grooves ("divots") during subsequent processing. Polysilicon 140 will be used to form the peripheral transistor gates. The grooving in regions 120i under the transistor gates is undesirable because it degrades the  
30 transistor characteristics.

[0044] Layer 140 can also be formed by non-conformal deposition processes,

whether known or to be invented. If the top surface of polysilicon 140 is not planar, it is believed that the polysilicon 140 can be planarized using known techniques (e.g. CMP, or spinning a photoresist layer over the polysilicon 140 and then simultaneously etching the resist and the polysilicon at equal etch rates until all of the photoresist is removed). The  
5 bottom surface of polysilicon 140 is non-planar as it goes up and down over the oxide protrusions 220P.

[0045] An exemplary final thickness of polysilicon 140 is 0.16  $\mu\text{m}$  over the active areas.

[0046] The peripheral area is masked, and polysilicon 140 is doped P+ in the array  
10 area. Polysilicon 140 remains undoped ("INTR", i.e. intrinsic) in the periphery. The peripheral transistor gates will be doped later, with the NMOS gates doped N+ and the PMOS gates P+, to fabricate surface channel transistors in the periphery with appropriate threshold voltages. The invention is not limited to the surface channel transistors or any peripheral processing. In particular, entire polysilicon 140 can be doped N+ or P+ after  
15 the deposition or in situ.

[0047] Silicon dioxide 810 is deposited on polysilicon 140, by CVD (TEOS) or some other process, to an exemplary thickness of 1500Å. Layer 810 can also be silicon nitride, silicon oxynitride (SiON), or some other material. Layer 810 is sufficiently thick to withstand subsequent oxide etches (and in particular the etch of STI oxide 220 described  
20 below in connection with Fig. 20A) and to protect the select gates 140 from counterdoping during subsequent doping steps.

[0048] In some embodiments, the top surface of polysilicon 140 and/or oxide 810 is not planar.

[0049] The wafer is coated with a photoresist layer 820. See Fig. 14A, cross section  
25 X1-X1', and Fig. 14B, periphery. (Fig. 14B shows only the active area, not the field oxide 220.) Resist 820 is patterned to define the select gate lines 140. The peripheral area is covered by the resist. The memory array geometry is not sensitive to a misalignment between mask 820 and the mask defining the isolation trenches 220T (Figs. 10A, 10B) except possibly at the boundary of the memory array.

30 [0050] Silicon dioxide 810 is etched through the resist openings. The resist is removed, and polysilicon 140 is etched away where exposed by oxide 810. Then the

exposed oxide 130 is removed. (In an alternative embodiment, the resist 820 is removed after the etch of polysilicon 140 and/or oxide 130.) The select gate lines are formed as a result. Each select gate 140 will control the conductivity of the underlying portion of the cell's channel region in substrate 120. Fig. 15 is a perspective view of the resulting structure in the array area.

[0051] The etch of polysilicon 140 can be a perfectly anisotropic vertical etch. Alternatively, the etch can have a horizontal component to reduce the width  $L_s$  (Fig. 14A) of select gate lines 140 (the width  $L_s$  is the channel length of the select gate transistor). In one embodiment, a perfectly vertical etch is performed first to remove the exposed portions of layer 140, and then an isotropic etch is performed to reduce the width  $L_s$ .

[0052] In another embodiment, one or more etching steps are performed as described above to form the lines 140. Then the sidewalls of lines 140 are oxidized. Substrate 120 is also oxidized in this step. The select gate line width  $L_s$  is reduced as a result. Then the oxide is removed.

[0053] The width  $L_s$  can also be reduced by a horizontal etch of layer 810. E.g., if layer 810 is SiON, a dry etch having a horizontal component can be used to pattern this layer.

[0054] In another embodiment, the sidewalls of the select gate lines are reacted with some material other than oxygen, with a reaction product forming on the sidewalls. The reaction product is then removed.

[0055] The lines 140 can thus be more narrow than the minimal photolithographic line width. The memory packing density is therefore increased.

[0056] As shown in Fig. 16 (cross section X1-X1'), the structure is oxidized to grow silicon dioxide 150 on substrate 120 and the sidewall surfaces of polysilicon gates 140 in the array area. Oxide 150 will serve as tunnel oxide on substrate 120, and will provide sidewall insulation for the select gates. The oxide thickness depends on the dopants and dopant concentrations. In some embodiments, oxide 150 is 60 Å to 100 Å thick on substrate 120, and is 300 Å thick on the select gate sidewalls. The peripheral area is covered by oxide 810 (Fig. 13B), and remains substantially unchanged during this step.

Oxide 150 can be nitrided to prevent boron diffusion from floating gates 160 into substrate 120 if the floating gates will be doped with boron. In the embodiment being

described, the floating gates will be doped P+ to improve the data retention time. (The data retention is improved because the P+ doped polysilicon is a high work function material. See U.S. patent no. 6,518,618 issued February 11, 2003 to Fazio et al. and incorporated herein by reference.)

5 [0057] If desired, an additional Vt adjust implant can be performed into the array to adjust the threshold voltage of the floating gate transistors (FG/CG transistors). This implant can be performed either before or after the formation of oxide 150. In one embodiment, the implant is performed after the etch of polysilicon 140 to define the select gates (Fig. 14A) before the removal of oxide 130 from the FG/CG channel areas.  
10 The floating gate transistors can be either enhancement or depletion mode transistors.

[0058] Floating gate polysilicon 160 (Fig. 17, cross section X1-X1') is deposited over the structure, by LPCVD for example, and is doped P+ during or after the deposition. Polysilicon 160 is sufficiently thick to ensure that its top surface is at least as high throughout the wafer as the top surface of oxide 810. In the embodiment of Fig. 17, the  
15 top surface of layer 160 is planar due to a conformal deposition to a thickness larger than half the distance between the adjacent select gate lines 140. In one embodiment, the distance between select gate lines 140 is 0.8  $\mu\text{m}$ , and the polysilicon 160 is more than 0.4  $\mu\text{m}$  thick.

[0059] If the top surface of polysilicon 160 is not planar, it is planarized by CMP or a  
20 suitable etch.

[0060] After planarization (if needed), layer 160 is etched down without a mask. The etch end point is when STI oxide 220 becomes exposed. Fig. 18 (cross section X1-X1') shows an intermediate stage in this etch, when oxide 810 becomes exposed. At this stage, layer 160 has been removed from the periphery, so the periphery becomes as in Fig. 13B.  
25 The etch endpoint can be the exposure of oxide 220. The endpoint is well defined if the layer 810 is SiON or silicon nitride, but it is also possible to detect the exposure of oxide 220 if layer 810 is silicon dioxide. Alternatively, the etch can be programmed as a timed etch continuing for a predetermined time after the exposure of layer 810.

[0061] Figs. 19A (cross section X1-X1') and 19B (cross section Y2-Y2') show the  
30 array area at the end of the polysilicon etch. The polysilicon has been removed from the top surface of oxide 220. In some embodiments, the final thickness of layer 160 is 1200Å.

The etch is selective to oxide 810.

[0062] Optionally, a timed etch of oxide 220 is performed to recess the top surface of oxide 220 below the surface of polysilicon 160. See Fig. 20A (cross section Y2-Y2') and Fig. 20B (perspective view of the array). This etch will improve the capacitive coupling between the floating and control gates. See the aforementioned U.S. patent no. 6,355,524. In the embodiment of Figs. 20A, 20B, the oxide 220 continues to protrude above the top surface of substrate 120 by about 0.10  $\mu\text{m}$ . In other embodiments, the oxide 220 does not protrude above the substrate after the etch (the top surface of layer 220 is level with the top surface of the substrate after the oxide etch).

[0063] As mentioned above, layer 810 is sufficiently thick to withstand this etch.

[0064] ONO layer 164 (Fig. 21A, cross section X1-X1', and Fig. 21B, periphery) is formed over the structure. Control gate polysilicon layer 170 is deposited on ONO 164 and is doped during or after the deposition. This layer is doped N<sup>+</sup> in the embodiment being described, P<sup>+</sup> in other embodiments. This may also be a metal or metal silicide layer, or some other conductive material.

[0065] The top surface of polysilicon 170 is not planar in the array area. Layer 170 has protrusions 170.1 over the select gate lines 140. Cavities 170C form in layer 170 between protrusions 170.1 over the future positions of bitline regions 174. The protrusions 170.1 will be used to define the overlap between the floating and control gates without additional dependence on photolithographic alignment.

[0066] As shown in Fig. 22 (cross section X1-X1'), a layer 1710 is deposited over the structure and etched without a mask to expose the polysilicon 170. Layer 1710 fills the cavities 170C. When layer 1710 is etched in the array area, layer 1710 is removed in the periphery, so the periphery becomes as in Fig. 21B. In one embodiment, layer 1710 is silicon nitride deposited to have a planar top surface or planarized during the etch.

[0067] Polysilicon 170 is etched without a mask. See Fig. 23A (cross section X1-X1') and 23B (periphery). This etch attacks the polysilicon portions 170.1 and exposes ONO 164. Polysilicon layer 170 becomes broken over the select gate lines 140. In other words, the polysilicon etch creates a gap 170G (a through hole) in polysilicon layer 170 over each select gate line 140. In the embodiment of Fig. 23A, the etch endpoint is the exposure of ONO 164. In other embodiments, the etch continues after the

exposure of ONO 164. In either case, at the conclusion of the polysilicon etch, polysilicon 170 is exposed near the select gates 140 but some of polysilicon 170 is covered by nitride 1710. The width W1 of the exposed portions of polysilicon layer 170 adjacent to gaps 170G will define the width of the control and floating gates in a self-aligned manner as illustrated below.

[0068] In some embodiments, the minimum thickness of polysilicon 170 (near the gaps 170G) is 0.18  $\mu\text{m}$ , and the width W1 is also 0.18  $\mu\text{m}$ .

[0069] In the embodiment of Fig. 23A, the etch of polysilicon 170 is selective to nitride 1710. In other embodiments, the etch is not selective to the nitride, and nitride 1710 is etched at the same rate as the polysilicon. The etch can stop on the top oxide sub-layer of ONO 164. The etch can be replaced with CMP. In some embodiments, the etch or the CMP removes some or all of ONO 164 above the select gates 140 and exposes the oxide 810. In either case, at the conclusion of the etch or the CMP process, polysilicon 170 is exposed near the select gates 140 but some of polysilicon 170 is covered by nitride 1710. The width W1 of the exposed polysilicon portions will define the width of the control and floating gates as illustrated below.

[0070] A protective layer 1910 (Fig. 24, cross section X1-X1') is formed adjacent to gaps 170G to protect the polysilicon 170 near the select gates 140. In one embodiment, layer 1910 is silicon dioxide formed by thermal oxidation of layer 170. An exemplary thickness of oxide 1910 is 500Å. Layer 1910 can also be a conductive metal silicide formed selectively on polysilicon 170 by a silicide (self-aligned silicidation) technique. In another embodiment, layer 1910 is deposited over the whole wafer and then removed by CMP from the top surface of layer 1710. See U.S. patent application no. 10/393,212 filed March 19, 2003 by Yi Ding and incorporated herein by reference.

[0071] Nitride 1710 is removed (by a wet etch for example) selectively to oxide 1910. The resulting structure is shown in Fig. 25 (cross section X1-X1'). The periphery remains as in Fig. 23B.

[0072] Polysilicon 170, ONO 164, and polysilicon 160 are etched with oxide 1910 as a mask. The resulting structure is shown in Fig. 26A (cross section X1-X1') and Fig. 26B (periphery). In some embodiments, the polysilicon etch of layers 170, 160 is anisotropic, and the etch of ONO 164 is isotropic or anisotropic. The ONO etch may remove the ONO

164 over the select gates 140 and may also remove portions of oxide 1910 and/or oxide 810.

[0073] In each FG/CG stack, the floating gate 160 together with control gate 170 control the underlying portion of the cell's channel region.

5 [0074] A photoresist layer (not shown) is formed over the wafer and patterned to cover the array but expose the entire periphery. Then oxide 810 (Fig. 26B) is etched away from the peripheral area.

[0075] The resist covering the array is removed, and another photoresist layer (not shown) is formed to cover the array and define the peripheral transistor gates. Polysilicon  
10 140 is etched away where exposed by this resist.

[0076] The resist is removed. The wafer is coated with a photoresist layer 2720 (Fig. 27B, periphery). The resist is patterned to expose the entire array area (Fig. 27A, cross section X1-X1') and also to expose the peripheral NMOS transistor regions. Fig. 27B shows a peripheral NMOS transistor region 512N with a P well 2724P, and a peripheral  
15 PMOS transistor region 512P with an N well 2724N. These wells were defined before formation of oxide 130. There can be many regions 512N, 512P in the integrated circuit. Resist 2720 covers the PMOS transistor regions 512P. An N type implant (N-) is performed to form the LDD (lightly doped drain) extensions for peripheral NMOS source/drain regions 2730N (Fig. 27B). This implant also dopes the NMOS gates 140 in  
20 the periphery. In addition, the implant dopes bitline regions 174 (Fig. 27A).

[0077] In some embodiments, the memory array is not exposed by resist 2720, and no doping is performed in the bitline regions at this step.

[0078] Resist 2720 is removed, and another photoresist layer 2820 (Fig. 28, periphery) is formed to cover the NMOS peripheral transistor regions 512N and the  
25 memory array. A P type implant (P-) is performed to form the LDD extensions for PMOS source/drain regions 2730P and to dope the peripheral PMOS transistor gates.

[0079] Resist 2820 is removed. A thin silicon dioxide layer 2904 (see Fig. 29A, cross section X1-X1', and Fig. 29B, periphery) is grown on the exposed silicon surfaces of layers 140, 160, 170 by a rapid thermal oxidation process (RTO). Alternative techniques  
30 can also be used such as chemical vapor deposition (e.g. TEOS CVD), a high temperature

oxide process (HTO), or other suitable techniques, known or to be invented. These techniques may form the oxide 2904 over the entire structure and not only on the silicon surfaces. An exemplary thickness of oxide 2904 is 100 Å.

5     **[0080]**     A silicon nitride layer 2910 is deposited to an exemplary thickness of 500 Å to 800 Å. Layer 2910 is etched anisotropically without a mask to form sidewall spacers over the gate structures. The etch of nitride 2910 may remove some of oxide 810 in the array area (Fig. 29A). If oxide 2904 was deposited over the entire structure (by TEOS CVD or HTO for example), oxide 2904 will help protect the substrate 120 during the nitride etch.

10     **[0081]**     Then N<sup>+</sup> and P<sup>+</sup> implants are performed to create source/drain structures for the peripheral transistors and the bitline regions 174. More particularly, the peripheral PMOS transistor area 512P is masked with resist (not shown), and an N<sup>+</sup> implant is performed to create the source/drain structures for bitline regions 174 and the peripheral NMOS transistors and increase the dopant concentration in the peripheral NMOS gates 140. The floating, control and select gates and the overlying layers mask this implant so  
15     no additional masking in the array area is needed.

**[0082]**     The resist is removed. The array and the peripheral NMOS transistor regions 512N are masked with a resist (not shown), and a P<sup>+</sup> implant is performed to create the source/drain structures for the peripheral PMOS transistors and increase the dopant concentration in the PMOS transistor gates 140.

20     **[0083]**     The resist is removed. A silicon dioxide etch is performed to remove the oxide 1910 and expose the control gate lines 170 (Fig. 30A, cross section X1-X1'). This etch also removes the exposed portions of oxide 150 over bitline regions 174 in the array area, the exposed oxide 130 over source/drain regions 2730N, 2730P in the periphery (see Fig. 30B), and the oxide 2904 over the peripheral transistor gates.

25     **[0084]**     A conductive metal silicide layer 2920 is formed by a self-aligned silicidation (salicide) process on the exposed silicon surfaces of control gate lines 170, bitline regions 174, peripheral transistor gates 140 and peripheral source/drain regions 2730N, 2730P. The salicide process involves depositing a metal layer, heating the structure to react the metal with the silicon, and removing the unreacted metal. This can be followed by an  
30     anneal or any other suitable processing, known or to be invented, to improve the silicide properties (e.g. increase its conductivity). Titanium, cobalt, nickel, and other conductive

materials, known or to be invented, can be used for the metal layer. Non-salicide selective deposition techniques, known or to be invented, that selectively form a conductive layer 2920 on the exposed silicon but not on a non-silicon surface, can also be used. Silicide 2920 has a lower resistivity and a lower sheet resistance than polysilicon 170.

5   **[0085]**     As noted above in connection with Fig. 24, layer 1910 can be a conductive metal silicide formed by a salicide process. In this case, layer 1910 does not have to be removed. The silicidation process of Fig. 30A will silicide the bitline regions 174, the peripheral gates 140 and the peripheral source/drain regions 2730.

10   **[0086]**     As shown in Fig. 31A (cross section X1-X1'), Fig. 31B (array boundary or an array gap without floating gates), and Figs. 31C and 31D (periphery), inter-level dielectric 3204 is deposited over the wafer. Fig. 31C shows only an NMOS transistor region, but the PMOS regions are similar. Contact openings are etched in dielectric 3204 to expose the silicided surfaces of bitline regions 174 (Fig. 31A), control gates 170 (Fig. 31B), peripheral source/drain regions 2730N and 2730P (Figs. 30B, 31C), and peripheral  
15   gates 140 (Fig. 31D). The silicide 2920 protects the bitline regions 174 and the source/drain regions 2730 during this etch. A conductive layer 3210 (e.g. metal) is deposited and patterned to contact the silicided regions. The figures also show an optional metal layer 3220 (e.g. tungsten) used to fill the contact openings before the deposition of layer 3210.

20   **[0087]**     In the embodiment of Fig. 31A, metal 3210 is used to form jumpers between the adjacent bitline regions 174 connected together (see Fig. 9). Then another dielectric layer 3230 (not shown in Figs. 31B-31D) is deposited, contact openings are etched in this layer to jumpers 3210, and another metal layer 3240 is deposited on top and patterned to form the bitlines 180. The bitlines contact the bitline regions 174 through the jumpers  
25   made from metal 3210. The openings in layer 3240 are filled with optional tungsten plugs 3250 before the metal 3240 is deposited.

30   **[0088]**     Fig. 31E (top view) shows an extension of a peripheral transistor gate 140 over STI oxide 220. The extension can be made to form a contact to the gate or for some other reason (e.g. to connect the gate to other features). The region 120i at the interface between the substrate 120 and field oxide 220 is protected from the divot formation because the gate is formed using the first polysilicon layer 140. See also Fig. 13B. The transistor of Fig. 31E can be a high voltage transistor (in area 512H in Fig. 12B) or a low

voltage transistor (in area 512L).

[0089] In Fig. 30A, the width of select gate 140 is shown as  $L_s$ , and the width of each of floating gates 160 is shown as  $L_f$ . The floating gate width  $L_f$  is defined by the parameter  $W1$  (Fig. 23A) in a self-aligned manner, so  $L_f$  can be smaller than the minimal photolithographic line width.  $L_s$  can also be smaller than the minimal photolithographic line width as explained above in connection with Fig. 14A.  $L_s$  can be smaller than  $L_f$ , or can be equal to or larger than  $L_f$ .

[0090] In each bit of the memory cell, ONO layer 164 forms a continuous feature overlying the respective floating gate and overlaying a sidewall of select gate line 140. This feature extends the whole length of the select gate line 140 (in the Y direction). Control gate 170 overlies the continuous feature of ONO 164. The portion of ONO 164 overlaying the sidewall of select gate line 140 separates the control gate 170 from the select gate 140.

[0091] Other details of the memory fabrication process for one embodiment are given in U.S. patent application no. 10/393,212 "NONVOLATILE MEMORIES AND METHODS OF FABRICATION" filed March 19, 2003 by Yi Ding and incorporated herein by reference.

[0092] Fig. 32 shows an alternative layout of the array. Here the connection between the source/drain regions 174 in the adjacent columns is done through the substrate 120. Each contiguous  $N^+$  type region 174 provides two source/drain regions for one of the two adjacent columns and also provides two source/drain regions 174 for the other one of the adjacent columns. In the first and last rows of the array, each region 174 provides one source/drain region for each of the two adjacent columns. Jumpers made from layer 3210 of Fig. 31A are unnecessary. Layer 3210 can be used to form the bitlines 180. The number of bitline contact openings 174C can be reduced, because only one contact is needed for each pair of source/drain regions 174 that are shorted together. Other layouts are also possible.

[0093] In some embodiments, the memory cells are read, programmed and erased using the same voltages and mechanisms as the cell of Fig. 5. The programming is done by channel hot electro ejection (CHIE) or Fowler-Nordheim tunneling. The voltages can be as in Figs. 6-8. Other exemplary voltages are shown in the following Table 1:

[0094]

TABLE 1

	Read	Program (CHEI)	Erase
<u>Select gate 140</u>			
Selected row:	2.5V	1.5V	2V
Unselected row:	0V	0V	0V
<u>Control gate 170</u>			
Selected row:			
Selected bit (Left or Right):	1.5V to 2V	9V to 10V	-9V to -10V
Unselected bit:	7V to 7.5V	7V to 7.5V	0V
Unselected row:	0V	0V	0V
<u>Bitline 180</u>			
Selected column:			
Selected bit:	1.5V	4.5V to 5V	Floating
Unselected bit:	0V	0V	0V
Unselected column:	0V	0V	0V
Substrate 120:	0V	0V	7V to 8V

[0095] The erase operation is through the channel region in substrate 120 (bulk erase). In other embodiments, the memory is erased through a source/drain region 174.

- 5 The programming can be performed by Fowler-Nordheim tunneling. In some embodiments, the programming is performed by an electron transfer between floating gate 160 and select gate 140.

- [0096] The invention is not limited to any particular read, erase or programming techniques, or to particular voltages. For example, the memory can be powered by multiple power supply voltages. Floating gates 160 can be defined using a masked etch, and can extend over sidewalls of select gate lines 140. See U.S. patent application no. 10/411,813 filed by Yi Ding on April 10, 2003 and incorporated herein by reference. Select gates 140 and/or floating gates 160 may be doped N+, and/or may include non-
- 10

semiconductor materials (e.g. metal silicide). The invention is not limited to the arrays of Fig. 9. Also, substrate isolation regions 220 do not have to traverse the entire array. The invention is applicable to non-flash memories (e.g. non-flash EEPROMs) and to multi-level memory cells (such a cell can store multiple bits of information in each floating gate). Other embodiments and variations are within the scope of the invention, as defined by the appended claims.